Exercices: distributions discrètes

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2018-11-19

Exercice 04.1 : probabilité d'un motif avec erreurs

On recherche dans un génome les occurrences du motif GATAAG en admettant un certain nombre de substitutions. En supposant que les nucléotides sont indépendants et équiprobables, quelle est la probabilité de trouver à une position du génome:

a. Une instance exacte du motif (aucune substitution) ?
b. Une séquence ne présentant aucune correspondance avec le motif (6 substitutions) ?
c. Une instance avec exactement 1 substitution ?
d. Une instance avec au plus 2 substitutions ?

Exercice 04.2 : alignement de lectures NGS

Au terme d'un séquençage de type *Next Generation Sequencing* (*NGS*), on dispose d'une librairie de $N = 10^6$ lectures courtes. On l'aligne sur un génome de référence, dont la somme des chromosomes fait $G = 10^9$ paires de bases, en utilisant un algorithme d'alignement sans gap et sans admettre aucune substitition.

On voudrait calculer de probabilité d'un alignement parfait (sans erreur) entre une séquence de lecture particulière à une position particulière du génome, en fonction de la longueur de lecture (k).

- Quelle distribution théorique utiliseriez-vous pour modéliser ce problème ? Justifiez ce choix.
- Ecrivez la formule de la probabilité.

Note: durant les travaux pratiques, nous dessinerons cette distribution avec le logiciel *R*.

Exercice 04.3 : sites de restriction

Dans un génome bactérien de 4 Mb avec une composition de 50% de G+C, on observe 130 occurrences de l'hexanucléotide GGCGCC. On suppose un schéma de Bernoulli et une composition équiprobable de nucléotides.

(a) Quelle est la probabilité d'observer une occurrence de GGCGCC à une position donnée du génome ?
(b) Combien d'occurrences s'attend-on à trouver dans l'ensemble du génome ?
(c) Quelle serait la probabilité d'observer un nombre aussi faible d'occurrences (130 ou moins) si l'on générait une séquence aléatoire selon le modèle de Bernoulli avec nucléotides équiprobables ?
(d) Comment peut-on interpréter cette sous-représentation de l'hexanucléotide GGCGCC du point de vue biologique ?

Exercice 04.4 : Jeu de roulette

La roulette comporte 37 nombres allant du 0 au 36. Un joueur a décidé de miser systématiquement 1 euro sur le nombre 17 jusqu'à ce que ce nombre sorte, et de s'arrêter ensuite.

Sachant que quand on mise sur un seul nombre, le gain vaut 36 fois la mise, quelle est la probabilité pour que le joueur sorte du casino en ayant gagné de l'argent ?

Il n'est pas nécessaire de fournir une réponse numérique, vous pouvez vous contenter de fournir la formule, en indiquant les nombres correspondant aux différents symboles mathématiques. Justifiez votre réponse en expliquant votre raisonnement.

Exercice 04.5 : probabilité des longueurs d'ORF

On détecte les cadres ouverts de lecture (*open reading frames*, *ORF*) d'un génome en identifiant toutes les séquences de taille multiple de 3 comprises entre un start (ATG) et un stop (TAA, TAG ou TGA).

a. Sur base des fréquences génomiques de trinucléotides, calculer la probabilité de trouver à une position donnée du génome un ORF d'au moins 100 codons.
b. Sachant que le génome fait 12 Mb, quel est le nombre attendu d'ORF d'au moins 100 codons ?

sequence	frequency	occurrences
AAA	0.0394	478708
ATG	0.0183	221902
TAA	0.0224	272041
TAG	0.0129	156668
TGA	0.0201	244627

Exercice 04.6 : mutagenèse

On soumet une librairie de molécules d'ADN de 1 kilobase à un traitement mutagène qui provoque un nombre moyen de 5 mutations ponctuelles (substitutions) par molécule.

a. Quelle est la probabilité d'avoir exactement 5 mutations pour une molécule donnée ?
b. Quelle est la probabilité pour une molécule d'ADN de n'avoir subi aucune mutation au cours du traitement ?
c. Quelle est la probabilité d'obtenir au moins 10 mutations ?

Formulation attendue pour la réponse.

- Expliquez le raisonnement qui vous permet de modéliser ce problème.
- Justifiez vos choix des hypothèses de travail.
- Ecrivez les formules avec les symboles, puis remplacez le symboles par les valeurs numériques correspondantes. Il n'est pas nécessaire de calculer la valeur finale (ceci nécessite un ordinateur, nous le ferons pendant les TP).